

The 11th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 6-9, 2020, Warsaw, Poland

Diabetes Diagnostic Prediction Using Vector Support Machines

Amelec Vilorio^{a*}, Yaneth Herazo-Beltran^b, Danelys Cabrera^c, Omar Bonerge Pineda^d

^{a,c} Universidad de la Costa, Barranquilla, Colombia

^b Universidad Simon Bolivar, Barranquilla, Colombia

^d Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras

Abstract

The most important factors for the diagnosis of diabetes mellitus (DM) are age, body mass index (BMI) and blood glucose concentration. Diagnosis of DM by a doctor is complicated, because several factors are involved in the disease, and the diagnosis is subject to human error. A blood test does not provide enough information to make a correct diagnosis of the disease. A vector support machine (SVM) was implemented to predict the diagnosis of DM based on the factors mentioned in patients. The classes of the output variable are three: without diabetes, with a predisposition to diabetes and with diabetes. An SVM was obtained with an accuracy of 99.2% with Colombian patients and an accuracy of 65.6% with a data set of patients of a different ethnic background.

© 2020 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Medical Diagnosis, Diabetes Mellitus, Medical Computing, Machine Learning, Vector Support Machines.

1. Introduction

Diabetes mellitus (DM), by definition of the World Health Organization (WHO), is a chronic-degenerative disease caused by insufficient insulin production in the pancreas or by the body's inability to effectively use insulin produced, taking hyperglycemia (increased blood glucose) as the main indicator.

* Corresponding author. Tel.: +57-3046238313.

E-mail address: aviloria7@cuc.edu.co

In its early stage, DM usually produces no noticeable symptoms, but when detected late and not receiving adequate treatment can lead to serious health complications, such as: heart attack, blindness, kidney failure, limb amputation and even premature death, the latter represents a decrease in life expectancy of between 5 and 10 years from the healthy average [1, 2]. An early diagnosis of the disease can significantly increase the patient's quality of life [3].

Diagnosis of DM is complicated because it is a multifactorial disease. To make a diagnosis, the physician should evaluate the results of a patient test and compare them with those of patients under similar conditions to analyze previous decisions [4]. Analysis of the factors influencing the diagnosis can be affected by human error as it is subject to the doctor's interpretation. Another important issue is that undiagnosed patients cannot be treated, so their quality of life can worsen considerably [5].

A blood test alone is not enough to make a correct diagnosis, as it is not discriminatory enough, and its interpretation may differ between populations with different characteristics. Diagnosis of DM is even more difficult due to the lack of a reliable, low-cost, high-performing test that can be universally applicable (or in the Mexican population), and the low capacity of health systems to identify and manage new cases of abets, especially in developing countries such as Colombia.

This highlights an advantage of machine learning over human capacity in the topic of medical diagnosis. Recently we have started talking about terms like medical mining and medical computing to refer to computer applications in medicine. These areas make use of computational tools for medical data processing to facilitate their interpretation [6], [7].

According to reference [8], it is not necessary to take into account many parameters to carry out a medical diagnosis of the DM, which would only unnecessarily increase the difficulty of prediction, since it is possible to carry out the prediction from 5 parameters which are measurable and not subject to human interpretation or patient bias. Age and glucose concentration in the blood plasma are crucial for proper identification of the disease.

Vector Support Machines (SVMs) are a set of computational algorithms capable of identifying and representing nonlinear relationships in complex systems [9]. SVMs have performed effectively in both regression and classification problems.

Several techniques have been used to make a prediction in the diagnosis of diabetes in patients [10], [11], [12]. However, no references to machine learning work applied to DM prediction were found in Mexican patients. SVMs have been validated in previous work as an effective algorithm for prediction in medical diagnostics [13], [14]. Obtaining a prediction of medical diagnosis of DM in patients allows early care to disease control, in addition to reducing diagnostic times and representing economic savings for the health system and the patient. In this work, it was proposed to use the well-known SVM algorithm to analyze a data set based only on measurable patient variables to carry out a prediction, following the proposal of simplicity of the inputs of [6].

2. Methodology

Body mass index (BMI), age, blood glucose concentration (CG) and prior medical diagnosis of DM (no diabetes, predisposition to diabetes and diabetes) of 500 patients from a public hospital in Colombia (for confidentiality reasons) were taken the name will not be provided in this paper). Figure 1 and 2 show the relationship between the diagnosis of patients with glucose level (1a) and BMI (1b) respectively. 80% of this dataset was used to train a nonlinear SVM classifier to predict DM diagnosis in new patients and the remaining 20% for validation [15].

Age, BMI and blood glucose were set as indicators and therefore inputs to the SVM, while diagnosis is the variable to be predicted (classified). The kernel used for both training and prediction was radial-based [16].

The 10-fold cross-validation method was used to validate the computational model. Accuracy, sensitivity, specificity, positive and negative prediction values and confusion matrix, commonly used parameters in medical diagnostic prediction, were used as SVM performance metrics.

Each patient's diagnosis is developed in a set according to the following criteria established by WHO; $CG < 99$ corresponds to a patient without diabetes, $99 < CG < 127$ is a patient with possible diabetes and $CG \geq 127$ a patient without diabetes.

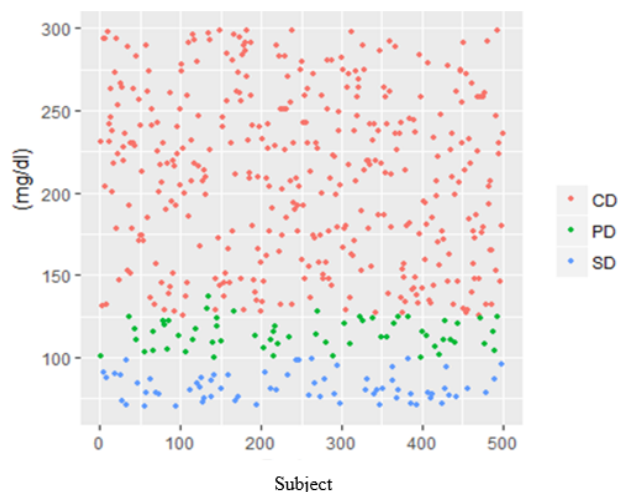


Fig. 1. Relationship between diagnosis and blood glucose level.

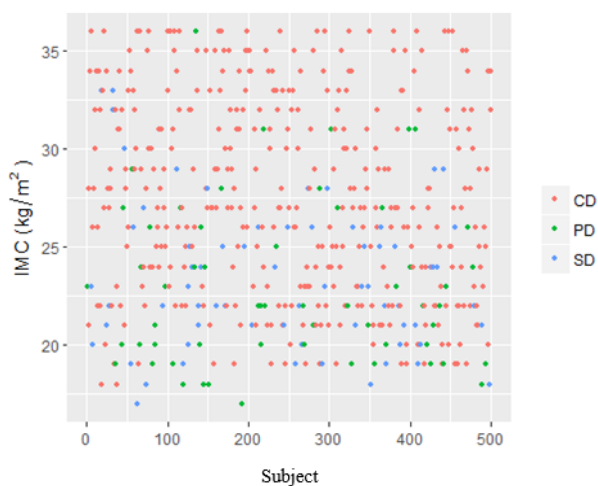


Fig. 2. Relationship between diagnosis and value IMC

3. Results

Table 1 shows the performance metrics for the model. Table II shows performance metrics in model validation with Pima Indians data [17]. Table III shows the actual diagnosis against prediction. Figure 3 shows the confusion matrix for the SVM-based model.

4. Discussion

A strong linear correlation is seen between GC in blood plasma and diagnosis of DM (Fig. 1(a)). An SVM was obtained with an accuracy of 95.36 %, which represents an acceptable value to use this technique in the diagnosis of DM in patients from Colombia with the ability to be applied in hospital patients across the country, improving the process of detecting to illness quickly, economically and correctly.

Table 1. model performance metrics

Metric	Percentage (%)
Accuracy	95.36
Sensitivity	94.36
Specificity	95.32
Positive prediction value	94.36
Negative prediction value	93.89

Table 2. Accuracy for the validation test with a second set of data

Metric	Percentage (%)
Accuracy	66.25
Sensitivity	45.99
Specificity	77.25
Positive prediction value	48.25
Negative prediction value	80.01

Table 3. Accuracy for the validation test with a third set of data

Real	Prediction
CD	CD
CD	PD
CD	PD
CD	CD
CD	CD
CD	CD

When testing the SVM in a different dataset than that used for training, taking into account the different characteristics of the diagnosed population (Table II). Acceptable accuracy (66.25%) was achieved, which validates the model developed even if there are differences between patients in both datasets, mainly ethnic in nature. It should be noted that the Pima Indians dataset takes as a criterion a glucose concentration greater than 250 mg/dL (after 2 hours after the intake of a food) to group the patient into the group diagnosed with diabetes, compared to criteria used for this work, the difference between the two is significant and therefore the accuracy is affected when validating the model [18].

Although the actual diagnosis corresponds to patients with diabetes, the model assigns the third set of data with predisposition to diabetes (Table 3), this because the blood glucose concentrations of these patients are at the threshold in the criteria to be assigned between groups. This confirms model performance metrics.

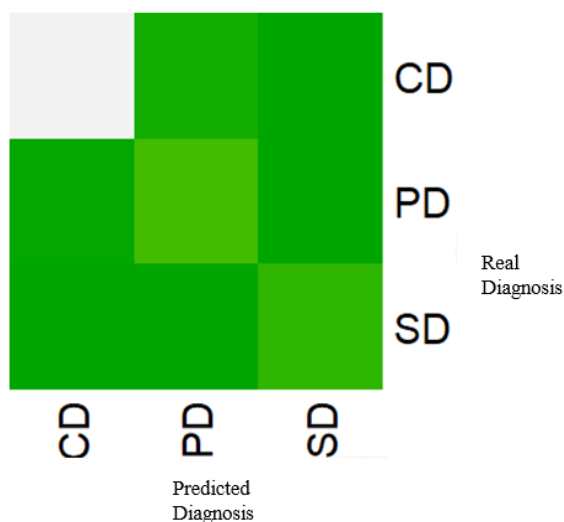


Fig. 3. Confusion matrix for the model built with the Support Vector Machine algorithm. SD: no diabetes, PD: diabetes predisposition, and CD: with diabetes.

5. Conclusions

An effective diagnostic dYG classifier based on the patient's age, BMI and CG was obtained. This classifier is a potential tool to help achieve good control over new DM cases in Colombia, as well as being an economical and universally applicable tool. New data and related attributes cCD on diagnosis of DM are necessary to test and improve this technique.

As future work, it is possible to increase the accuracy and predictability of the classifier using different algorithms, or by combining these with other computational techniques such as genetic algorithms or particle swarm optimization. In addition to this, the level of accuracy can be increased by incorporating other parameters that contribute to a correct diagnosis, such as the concentration of glycosylated hemoglobin A, a biological marker of high importance, which also gives an indication of the quality of the care the patient has to control their illness and health status [19].

References

- [1] Bates, D., Mäechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi: 10.18637/jss.v067.i01.
- [2] INEGI, “Estadística a Propósito del Día Mundial de la Diabetes,” *Día Mund. la Diabetes.*, p. 18, 2013.
- [3] T. Santhanam and M. S. Padmavathi, “Application of K-Means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis,” *Procedia Comput. Sci.*, vol. 47, no. C, pp. 76–83, 2014.
- [4] Viloria A., Lis-Gutiérrez JP., Gaitán-Angulo M., Godoy A.R.M., Moreno G.C., Kamatkar S.J. (2018) Methodology for the Design of a Student Pattern Recognition Tool to Facilitate the Teaching - Learning Process Through Knowledge Data Discovery (Big Data). In: Tan Y., Shi Y., Tang Q. (eds) *Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science*, vol 10943. Springer, Cham
- [5] S. Li, H. Zhao, Z. Ru, and Q. Sun, “Probabilistic back analysis based on Bayesian and multi-output support vector machine for a high cut rock slope,” *Eng. Geol.*, vol. 203, pp. 178–190, 2016.
- [6] T. Zheng et al., “A machine learning-based framework to identify type 2 diabetes through electronic health records,” *Int. J. Med. Inform.*, vol. 97, pp. 120–127, 2017.
- [7] Shankaracharya, D. Odedra, S. Samanta, and A. S. Vidyarthi, “Computational intelligence in early diabetes diagnosis: A review,” *Rev. Diabet. Stud.*, vol. 7, no. 4, pp. 252–261, 2010.
- [8] K. V. S. R. P. Varma, A. A. Rao, T. Sita Maha Lakshmi, and P. V. Nageswara Rao, “A computational intelligence approach for a better diagnosis of diabetic patients,” *Comput. Electr. Eng.*, vol. 40, no. 5, pp. 1758–1765, 2014.

- [9] D. Çalışır and E. Dogantekin, “An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier,” *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8311–8315, 2011.
- [10] H. Temurtas, N. Yumusak, and F. Temurtas, “A comparative study on diabetes disease diagnosis using neural networks,” *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8610–8615, 2009.
- [11] Mellado A., Suárez, N., Altimir, C., Martínez, C., Pérez J. C., Krause, M., & Horvath, A. (2017) Disentangling the change-alliance relationship: Observational assessment of the therapeutic alliance during change and stuck episodes. *Psychotherapy Research*, 27(5), 595-607. doi: 10.1080/10503307.2016.1147657
- [12] Ogles, B. M. (2013). Measuring change in psychotherapy research. En M. J. Lambert (Ed.), *Bergin and Garfields's Handbook of Psychotherapy and Behavior Change* (pp.134– 166). New Jersey: Wiley.
- [13] El Pasante, «Ventajas y desventajas de las bases de datos,» 17 Junio 2015. [En línea]. Available: <https://educacion.elpensante.com/ventajas-y-desventajas-de-las-bases-de-datos/>. [Último acceso: 12 Noviembre 2018].
- [14] Probability Formula, «Hypergeometric Distribution,» [En línea]. Available: <http://www.probabilityformula.org/hypergeometric-distribution.html>. [Último acceso: 16 Noviembre 2018].
- [15] Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. Boca Raton: Chapman & Hall/CRC
- [16] J. Swamidass† y P. Baldi, «Mathematical Correction for Fingerprint Similarity Measures to Improve Chemical Retrieval,» *Journal of Chemical Information and Modeling*, vol. 47, n° 1, pp. 952-964, 2006.
- [17] S. Arif, J. Holliday y P. Willett, «Comparison of chemical similarity measures using different numbers of query structures,» *Journal of Information Science*, vol. 39, n° 1, pp. 1-8, 2013.
- [18] Bucci, N., Luna, M., Viloria, A., García, J. H., Parody, A., Varela, N., & López, L. A. B. [2018, June). Factor analysis of the psychosocial risk assessment instrument. In *International Conference on Data Mining and Big Data* [pp. 149-158]. Springer, Cham
- [19] Viloria, A., Bucci, N., Luna, M., Lis-Gutiérrez, J. P., Parody, A., Bent, D. E. S., & López, L. A. B. (2018, June). Determination of dimensionality of the psychosocial risk assessment of internal, individual, double presence and external factors in work environments. In *International Conference on Data Mining and Big Data* (pp. 304-313). Springer, Cham.